CONTENTS

## 1 Introduction

## 2 Questionnaire contents

## 3 Machine Learning Methodologies

## 4 chi-square test and feature selection

## 5 distribution difference between two sets

## 6 feature selection using genetic algorithm

## 7 Imbalanced data handling

## 8 graph method

## 9 Conclusion and future research

1 introduction

Prediction of the student learning outcomes has been a popular topic in the Educational Data Mining (EDM) field. With the help of state-of-art machine learning technique, we have more and more powerful algorithm to predict the students’ learning outcome.

In this research, I test various Machine learning methods on the data which comes from questionnaires finished by 320 CUHK science student, where 89 of them belongs to Engineering Department (taking course MATH1510) while 231 of them belongs to Science Department (taking course MATH1010).

This report tests different machine learning methods on the dataset, like tree-based algorithm with entropy criteria and GINI criteria (C4.5 and CART); Naïve Bayes; aggregation method like Random-Forest, Adaboost. Also, I test the effective of turn categorical variables into numerical variables using one hot encoding, and apply numerical classification algorithms like SVM, logistic regression and artificial neural network (ANN) to see their behavior.

Another problem we may consider is that the data is ill-skewed. For example, the number of data comes from Science student is 3 times compare with the number of data comes from Engineering students. And people may expect that the male student is more than female student, actually the gender ratio of Science students is 8:2. The method SMOTE is tried during the experiment to see whether the data imbalance is possible to be solved.

This report will analysis the distribution difference between students with different gender and Class. And it is an interesting finding that by separately training predictors using data from Science and Engineering, the precision is significantly better than combining data and using single predictor. In addition, graph theories like minimal spanning tree and centrality evaluation is applied to intuitively identify the distribution difference between students from two classes.

Some data selection methods, like Chi-square test and Genetic algorithm are applied to found out the most importance questions in the questionnaire to predict student learning outcome. And ways to handle the imbalanced data are discussed in the end.

2 Questionnaire contents

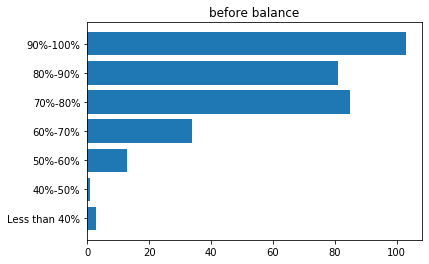
## 2.1 Questionnaire contents

There are 32 questions in the questionnaire, and all of them are of multiple choice form, which means that all variables in the data set is categorical variables.

1. Student’s gender:
2. Mother’s education:',
3. Mother’s occupation:
4. Father’s education:
5. Father’s occupation:
6. School type:
7. Reason you chose your former secondary school:
8. Weekly study time in MATH on average:
9. Join MATH discussion group regularly:
10. Work on MATH problems via Internet resources:
11. Raised hand in class:
12. Ask questions after class:
13. I write summary notes and I use them later when preparing for tests:
14. I schedule my study time carefully:
15. Getting a low test score does not make me feel sad:
16. Excited about the courses I take:
17. I control my upset mood or anger without blaming others:
18. I can make friends in new places:
19. I am open to people with different opinions:
20. I do outdoor/indoor exercises regularly:
21. My emotional health supports my ability to learn:
22. I plan for each day of the week:
23. I do not need to study hard the night before an exam because of good planning:
24. I see education as something I will be doing throughout my life:
25. I know that I am responsible for my own education:
26. I know a lot about myself in relation to what kind of profession would be best for me:
27. I know the education choices and schedules that I must follow in order to reach my career goal:
28. Available resources enough to meet my needs:
29. I do not hesitate to ask for help:
30. I am able to manage time and other life demands effectively:
31. I can do college-level work successfully:
32. Final overall grade achieved in your secondary school:

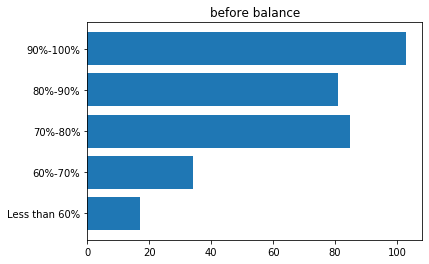
## 2.2 The distribution of target variables

For Machine Learning technique we will used latter, the target variable is the last question: “Final overall grade achieved in your secondary school:”. Its distribution is :



Since the cases: Less than 40%, 40%-50% and 50-60% are two small (only have 3,1,13 instance separately). Here I choose to combine them together as “Less than 60%”

Now the distribution becomes:



However, the distribution is still ill-skewed: the minority class “Less than 60%” has less than 20 records will the majority class “90%-100%” has more than 100 records. In other words, our data is imbalanced.

The major part of this report won’t take the data imbalance into consideration, and in chapter 7. We will come back and handling the problem.

3 Machine Learning Methodologies

## 3.1 Basic Machine Learning technique

Naïve Bayes classifier is a traditional multiclass classification algorithm. By computing the conditional probability distribution of each feature given label using training data, and applying Bayes’ theorem to compute the conditional probability distribution of target variable. Since it is simple and supported by classical statistic theorem, Naïve Bayes is very stable with almost 0 computation resources. However, since the Bayes’ theorem is built up under the assumption of independence between every pair of features, which is clearly not our cases, it is not surprising that Naïve Bayes is not the most suitable method for the prediction task.

Tree structure is a very common method for classification method, there are 2 main criterions for choosing the attributes which splitting nodes:

1. Entropy:

Entropy measures the pureness of our data, the formula is :

Where C is the number of levels and is the ratio of levels.

The tree algorithm tries to find out the attribute for each node which gives the smallest Entropy.

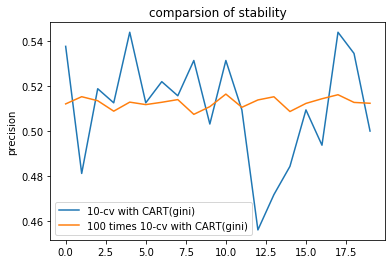
1. Gini index:

The Gini index is to measure how often a random element would be incorrectly identified, and also it can be regard as a metric which measures the pureness of data. The formula is:

Similar as entropy criterion, the tree algorithm tries to find out the attribute for each node which gives the smallest Gini index.

## 3.2 ways to make validation stable

Validation is always important for evaluation the power of particular method, and the common way for machine learning is cross-validation (10-folders-CV for example). However, even repeat the algorithm 10 times can’t give the stable result.



As the blue line shows, I did 20 times 10-folders-cv using CART (Gini criterion) with ‘rpart’ package. However, the precision varies from 0.45625 to 0.54375, which is unstable and result is misleading.

The orange line, whose values are calculated by 100 times 10-folder-cv, is much more stable, and the precision only varies from 50.74062 to 51.6375. It can be easily seen that repeat cross-validation is a good way to produce reliable evaluation.

In the rest of this report, unless pointed out specifically, all evaluations will use 10-folders-cv with 100 repeats.

## 3.3 One Hot Encoding

One hot encoding is a common way to transform categorical variable into numerical variable. For example, the question “Student’s gender” has two levels “male” and “female”, the One-Hot-Encoding will make the transformation:

male[1,0]

female[0,1]

By using one hot encoding, one benefit for our case this that we can apply numerical classification methods like logistic regression, SVM and ANN, which can not be used with categorical variables. In addition, we can also apply tree algorithm on the transformed numerical data, and surprisingly have better results than using the same method on categorical data.

For all the tree structure, I apply the pruning mechanism that limits the minimal node size larger or equals to 5.

For Gini tree using categorical variables, R package rpart is used.

For Entropy using categorical variables, R package Rweka (J48) is used.

For trees using numerical variables, python package sklearn is used.

|  |  |  |
| --- | --- | --- |
|  | Using categorical variables | Using transformed numerical variables |
| Naïve Bayes | 53.6875 |  |
| Gini tree | 51.15938 | 0.542875 |
| Entropy tree | 44.46563 | 0.543812 |

As we can see. By using numerical data (One-Hot-Encoding), the behavior of tree structures has obviously improvement. Depend on this inspiring finding, later in this report, all the numerical test and algorithm tried will be construct with One-Hot-Encoding.

## 3.4 aggregation method and classical numerical estimator

Random Forest is a common aggregation method that evolves from decision trees. By sampled data from the original dataset many times and take a randomly selected subset of features for each sample, many decision trees are constructed. With input data given, Random Forest will collects the classifications and chooses the most voted prediction as the result. Since there are many trees in a forest, the method is much more stable and insensitive for outlier, and pruning will not be applied for Random Forest.

In this report, I customize that there are always 200 trees in a forest. And make the criterion of trees Gini index.

Adaboost is another aggregation method which can not only be applied on tree structures, but can also be applied on almost arbitrary machine learning technique, such as logistic regression, SVM and even Random Forest.

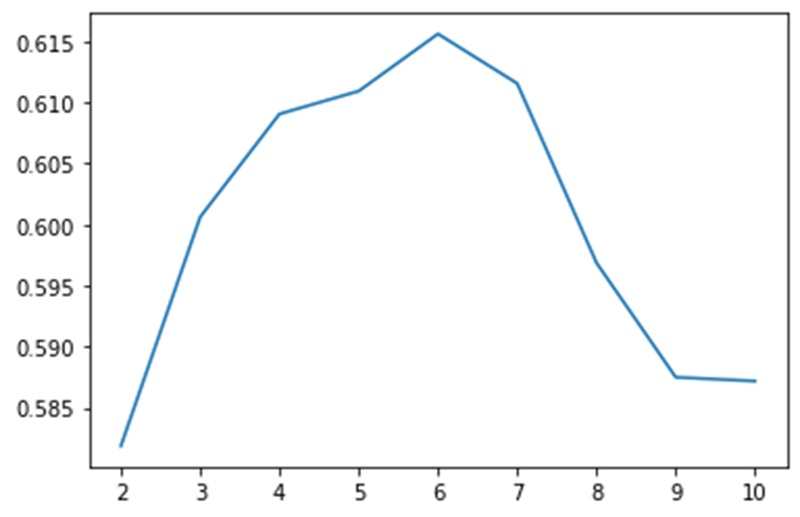
Unlike Random Forest, which contains many predictors trained on sampled data, and can be parallelly computed. The Adaboost also contains many predictors while each predictor is computed depend on its ex-predictor. The weight of each instances is different – if ex-predictor predicts something wrong, then the weight of wrong instances will increase and the weight of right case will decrease. Then a new predictor will be trained depend on the new weight.

In this paper, I set the predictor in each Adaboost algorithm 200 and the learning rate be 0.8

Also, I test the behavior of logistic regression and SVM. And choice of kernel of SVM is important, I do a simple numerical test to find the most suitable kernel for SVM:

|  |  |
| --- | --- |
| Gaussian kernel | 0.60625 |
| Linear kernel | 0.565625 |
| Ploy-2 kernel | 0.5875 |
| Ploy-3 kernel | 0.609375 |

The result suggests that poly kernel is the most suitable kernel for our case, and the highest precision it can achieve is 0.615625 with degree 6 poly-kernel.



Since Logistic regression and SVM are originally designed for binary classification task, while we are facing multi-classification task. There are two ways to solve this issue – one is one versus all (OVA) and another is Multinomial (only for logistic regression).

1. OVA:

A binary classification algorithm trained for each class, and distinguishes that class from all other classes. OVA is performed by look through all binary classification predictors and choose the most voted label.

1. Multinomial:

Unlike binary logistic regression, which calculated a score with sigmoid function

the multinomial method calculated SoftMax function for each level, and take the label with most confidence:

where C is the number of levels.

All models are trained and tested using python sklearn package.

The precision of different method are:

|  |  |
| --- | --- |
| method | precision |
| R]ndom forest | 0.6242187500000002 |
| Random forest + Adaboost | 0.62303125 |
| SVM (ploy-6 kernel) | 0.6144374999999997 |
| SVM (ploy-6 kernel) + Adaboost | 0.321875 |
| Tree (Gini) + Adaboost | 0.5739687499999999 |
| Tree (Entropy) + Adaboost | 0.5682187500000001 |
| Logistic+OVA+L1 penalty | 0.54375 |
| Logistic+OVA+L2 penalty | 0.56875 |
| Logistic+Multinomial+L1 penalty | 0.546875 |
| Logistic+Multinomial+L2 penalty | 0.575 |
| Logistic+OVA+L1 penalty + Adaboost | 0.321875 |
| Logistic+OVA+L2 penalty + Adaboost | 0.465625 |
| Logistic+Multinomial+L1 penalty + Adaboost | 0.321875 |
| Logistic+Multinomial+L2 penalty + Adaboost | 0.475 |

We can see that the Adaboost hardly improve the precision, and Random Forest have similar prediction accuracy compared with Random Forest + Adaboost and the best model can achieve 0.624 accuracy.

4 chi-square test and feature selection

There are 32 questions in the questionnaire, and to predict the student learning outcomes, the machine learning algorithms I used before considering all 31 variables. However, one natural question arises is that whether all these 31 variables are useful for predicting the student learning outcome? And there is a high probability that some features are not that important, and we may drop them.

One way to do feature selection is to directly consider the relationship between each feature and our target variable (student learning outcomes). Since we are dealing with categorical variables, chi-square test is the most commonly use way to test whether two variables are related.

Chi-squared test which is also called Pearson’s chi-squared test (χ2), is a statistical test applied on categorical data to evaluate the goodness of fit and independence.

The test-statistic is calculated by:

Where

χ2 = Pearson's cumulative test statistic, which asymptotically approaches a χ2 distribution.

= the number of observations of type i.

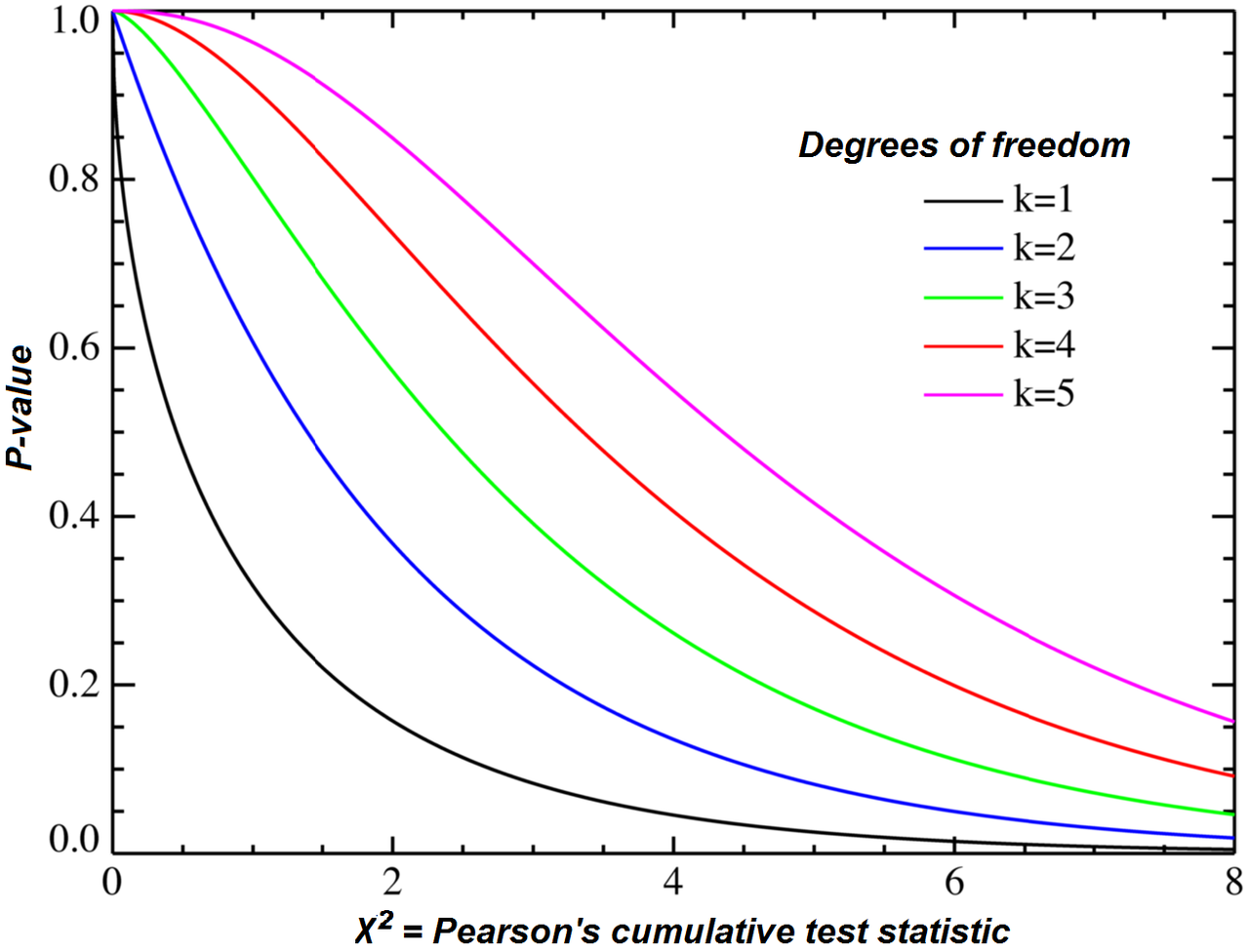
N = total number of observations

= null hypothesis that the fraction of type i in the population is

n = the number of cells in the contingency table.

Thus, under different freedom we can calculate the p-value from our test-statistic. The smaller the p-value is, the more confident we can conclude that there is relationship between two values.

And in this report, the chi-square test is constructed using R programing, and the Monte Carlo statistic test is used to generate p-values.



The Chi-square test have two usages:

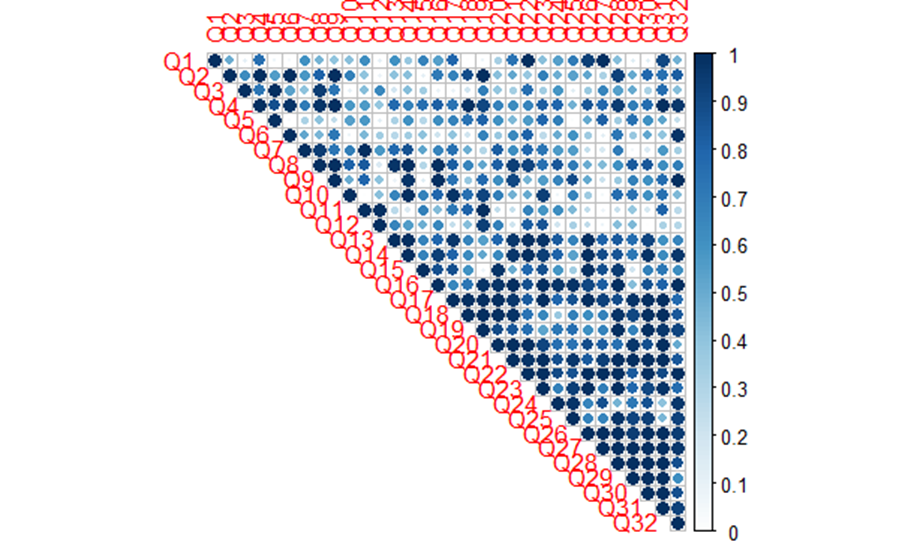
1. Relationship test:

Suppose two variables where (the number of levels of each variables), then the chi-square test is constructed on a  contingency table and the NULL hypothesis is no relationship between two variable.

1. Fitness test:

Suppose one variable but come from different set, where . The to test where are of same distribution, the chi-square test is constructed on a contingency tables and the NULL hypothesis is they have same distribution.

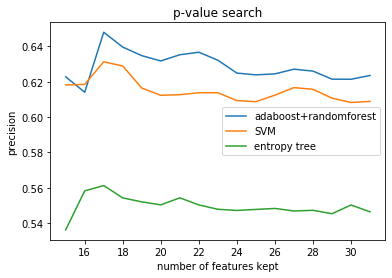
In this chapter, the relationship test is use and the chi-square table is plotted (deeper color means smaller p-value). In next chapter, where the difference between two set is discussed, fitness test will be applied.



The relationship between 31 features and target variable is:

|  |  |
| --- | --- |
| Question name | p-value related to target variable |
| Join MATH discussion group regularly | 4.9998e-05 |
| Excited about the courses I take | 4.9998e-05 |
| I plan for each day of the week | 4.9998e-05 |
| School type | 9.9995e-05 |
| Father’s education | 0.0001499925 |
| I schedule my study time carefully | 0.0001499925 |
| I see education as something I will be doing throughout my life | 0.0001499925 |
| I know the education choices and schedules that I must follow in order to reach my career goal | 0.0001499925 |
| I know a lot about myself in relation to what kind of profession would be best for me | 0.00019999 |
| I know that I am responsible for my own education | 0.0003499825 |
| I can do college level work successfully | 0.0003499825 |
| I am open to people with different opinions | 0.0004499775 |
| I am able to manage time and other life demands effectively | 0.00099995 |
| Available resources enough to meet my needs | 0.001099945 |
| My emotional health supports my ability to learn | 0.001649918 |
| I can make friends in new places | 0.00179991 |
| Mother s education | 0.002299885 |
| I control my upset mood or anger without blaming others | 0.00519974 |
| I do not need to study hard the night before an exam because of good planning | 0.005949703 |
| Weekly study time in MATH on average | 0.01554922 |
| I write summary notes and I use them later when preparing for tests | 0.02214889 |
| Getting a low-test score does not make me feel sad | 0.03349833 |
| I do not hesitate to ask for help | 0.03354832 |
| I do outdoor indoor exercises regularly | 0.07894605 |
| Student’s gender | 0.09884506 |
| Work on MATH problems via Internet resources | 0.1161442 |
| Mother’s occupation | 0.1348933 |
| Reason you chose your former secondary school | 0.2271886 |
| Ask questions after class | 0.2823359 |
| Raised hand in class | 0.3234338 |
| Father’s occupation | 0.368731 |

Now I do experiments to test what is the influence on the precision when the low-rank features (features with large p-value) is deleted:



We can see that there is a clearly increasing pattern with the number of features decreasing to 17. Which inspiring us that there are some useless variables and we should do feature selection to eliminate them. However, chi-square test is not the most suitable method to do feature selection, since:

1. It only tests the pairwise relationship while cannot put the interaction of features into consideration.
2. There are relationships between feature and target doesn’t means the feature is necessarily important for prediction target.

Thus, although by applying chi-square test and deleting features can improve the behavior of machine learning algorithms and improve the Random Forest + Adaboost method from 0.62303125 with 31 features to 0.64803125 with 17 features, the chi-square test is not the most desirable criterion for feature selection.

In chapter 6, we will use more suitable feature selection mechanism – genetic algorithm, which can take the interaction of features into consideration and has more globally view compared with chi-square test, to further explore the feature selection part.

5 distribution difference between two sets

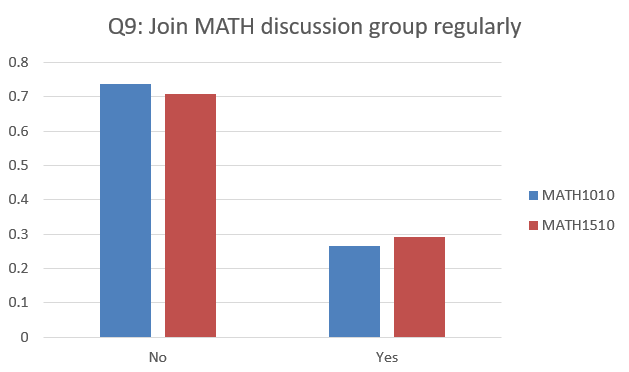
## 5.1 Distribution difference between data from different course

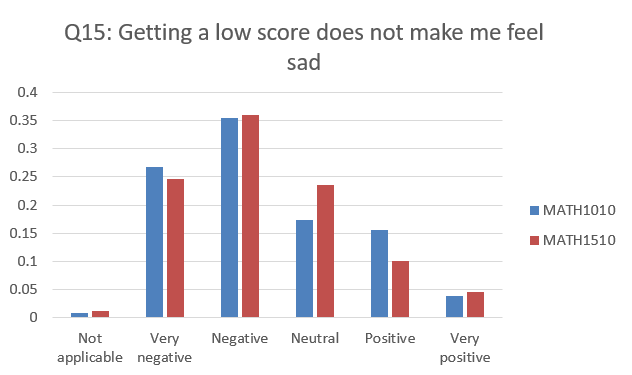
As the data we use comes from survey on students from two classes: MATH1010 (denoted as Class1) and MATH1510 (denoted as Class2). Whether the distribution of Class1 and Class2 is different is a important problem.

There I apply the chi-square fitness test to check whether there is any difference. Smaller p-value denoted larger probability that the distribution is different:

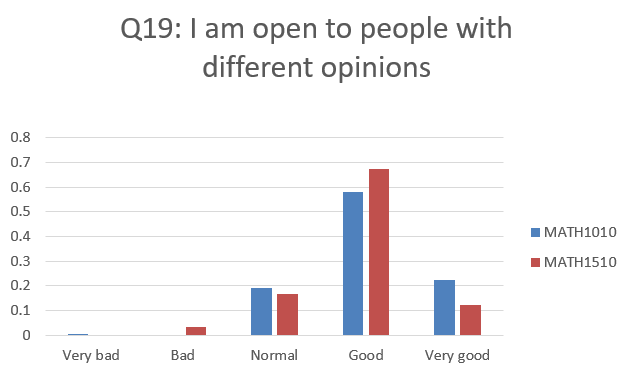
|  |  |
| --- | --- |
| Question content | p-value of fitness test |
| Join MATH discussion group regularly | 0.5481008 |
| Getting a low test score does not make me feel sad | 0.5435026 |
| I am able to manage time and other life demands effectively | 0.4663642 |
| Ask questions after class | 0.4464105 |
| My emotional health supports my ability to learn | 0.4189977 |
| Weekly study time in MATH on average | 0.3806002 |
| I know that I am responsible for my own education | 0.3074433 |
| School type | 0.2987119 |
| Father’s occupation | 0.2950974 |
| I can make friends in new places | 0.2823515 |
| I write summary notes and I use them later when preparing for tests | 0.2375082 |
| I control my upset mood or anger without blaming others | 0.1241305 |
| I know a lot about myself in relation to what kind of profession would be best for me | 0.09609377 |
| Available resources enough to meet my needs | 0.0802853 |
| Work on MATH problems via Internet resources | 0.07903179 |
| I see education as something I will be doing throughout my life | 0.07705866 |
| I do not need to study hard the night before an exam because of good planning | 0.06432221 |
| Mother’s occupation | 0.03259589 |
| Reason you chose your former secondary school | 0.02970153 |
| Excited about the courses I take | 0.02062835 |
| I do outdoor indoor exercises regularly | 0.01564111 |
| I know the education choices and schedules that I must follow in order to reach my career goal | 0.01463148 |
| I plan for each day of the week | 0.00995678 |
| I can do college level work successfully | 0.00854731 |
| Father’s education | 0.00225758 |
| I do not hesitate to ask for help | 0.0014403 |
| I schedule my study time carefully | 0.00106274 |
| Mother’s education | 0.00046586 |
| Raised hand in class | 0.000307 |
| Final overall grade achieved in your secondary school | 0.0001159 |
| I am open to people with different opinions | 8.4106e-05 |
| Student’s gender | 2.211e-08 |

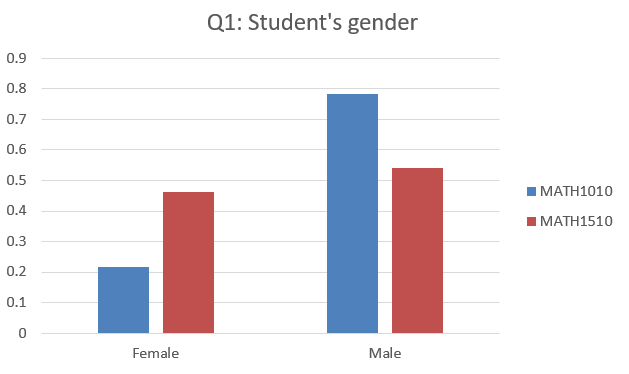
The distributions of some questions are very similar and have large p-values. For example, the questions “Join MATH discussion group regularly” and “Getting a low test score does not make me feel sad”, which have p-value 0.5481008 and 0.5435026 separately. As we may expected, the distribution of these two questions under two different set should be similar:



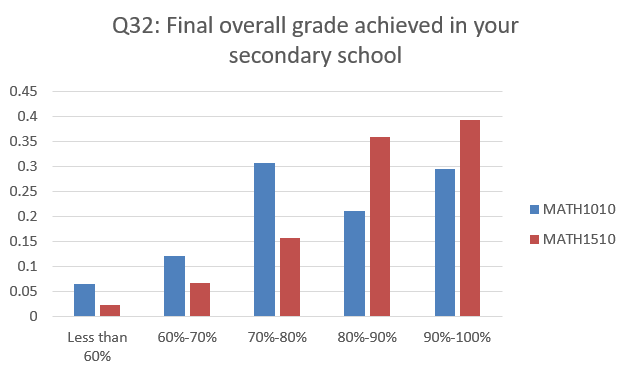


However, there are some questions whose answers have very different distribution under 2 sets. Like the questions “I am open to people with different opinions” and “Student’s gender”, whose fitness p-values are 8.4106e-05 and 2.211e-08 separately. We have nearly 100% confidence to conclude that their distributions are different between 2 sets.





Since our target is the question “Final overall grade achieved in your secondary school”, and its p-value of fitness test is 0.0001159, which is also very small. The distribution of this question is:



The distribution is quite different, which inspiring us to add a dummy variable “Class” into Machine learning model we have used.

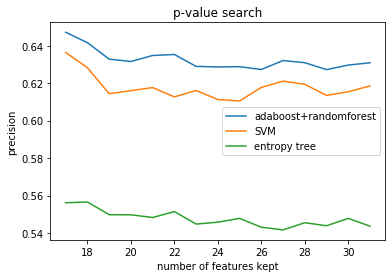
|  |  |  |
| --- | --- | --- |
| model | Without dummy variable | With dummy variable |
| SVM | 0.6143750000000001 | 0.6214687500000001 |
| Random Forest | 0.6228750000000003 | 0.63253125 |
| Random Forest + Adaboost | 0.6232187500000002 | 0.6313750000000001 |

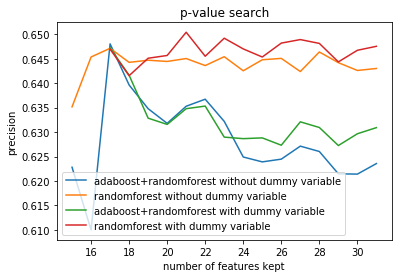
It is obvious that by adding dummy variable “Class”, all three techniques have improvement.

If we apply chi-square relationship test on the Class dummy variable, it’s p-value is 0.002549873. which is larger than the question “Mother’s education.” (rank 17th in chapter 4) and smaller than the question “I control my upset mood or anger without blaming others” (rank 18th in chapter 4). Then we do numerical test with iteratively eliminating low rank features.

The best precision achieved without dummy variable is Random Forest + Adaboost whose precision is 0.64803125 with 17 features

The best precision achieved with dummy variable is Random Forest whose precision is 0.65040625 with 21 features





The graph shown above suggests than using dummy variable can significantly improve the precision of Machine Learning technique.

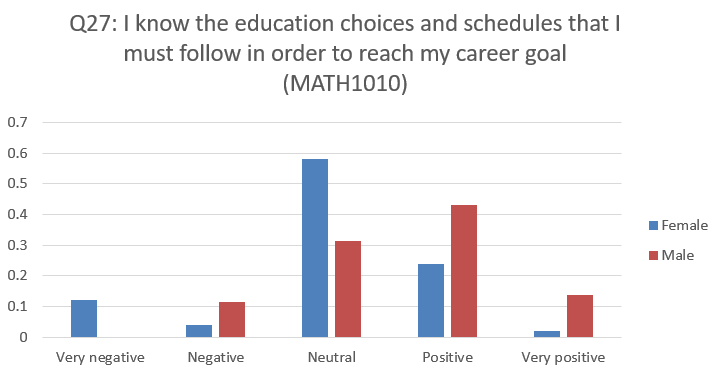
## 5.2 Distribution difference between student’s gender

There is no need to regard gender as a dummy variable, and the chi-square relationship test constructed in chapter 4 indicated that gender is not an important feature to predict the student learning outcomes. However, there is still worth learning whether there is any difference between students with difference gender.

Firstly, I separate the data from Class1 into Male group and Female group, and do the chi-square fitness test:

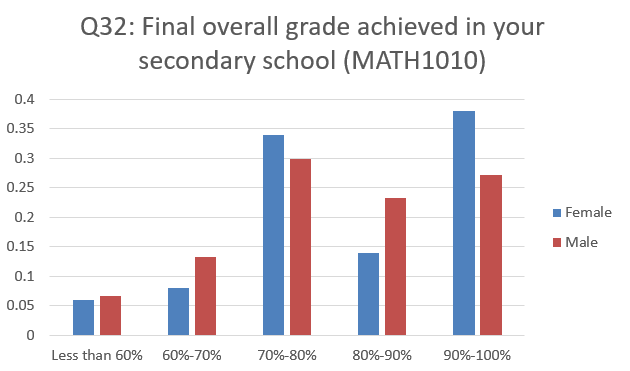
|  |  |
| --- | --- |
| Question content | p-value of fitness test |
| I can make friends in new places | 0.9366237 |
| I am open to people with different opinions | 0.7534646 |
| I schedule my study time carefully | 0.6457602 |
| I am able to manage time and other life demands effectively | 0.5570544 |
| I do not hesitate to ask for help | 0.3842717 |
| Father’s occupation | 0.3283092 |
| I see education as something I will be doing throughout my life | 0.301912 |
| I do not need to study hard the night before an exam because of good planning | 0.2936894 |
| Mother’s occupation | 0.2709673 |
| Raised hand in class | 0.2540768 |
| Final overall grade achieved in your secondary school | 0.2490993 |
| Join MATH discussion group regularly | 0.2428838 |
| I know that I am responsible for my own education | 0.1833796 |
| Mother’s education | 0.1658021 |
| Ask questions after class | 0.1577164 |
| Work on MATH problems via Internet resources | 0.1474195 |
| Father’s education | 0.1125734 |
| Reason you chose your former secondary school | 0.035727 |
| School type | 0.03040145 |
| Getting a low test score does not make me feel sad | 0.02271712 |
| I write summary notes and I use them later when preparing for tests | 0.0124027 |
| Excited about the courses I take | 0.007177587 |
| Weekly study time in MATH on average | 0.006401404 |
| I do outdoor indoor exercises regularly | 0.006342819 |
| Available resources enough to meet my needs | 0.0007074355 |
| I plan for each day of the week | 0.0001940738 |
| I know a lot about myself in relation to what kind of profession would be best for me | 1.473923e-05 |
| I control my upset mood or anger without blaming others | 5.092926e-08 |
| I can do college level work successfully | 7.963638e-13 |
| My emotional health supports my ability to learn | 2.342575e-13 |
| I know the education choices and schedules that I must follow in order to reach my career goal | 9.060143e-30 |

The question “I know the education choices and schedules that I must follow in order to reach my career goal” has a really small p-value for chi-square fitness test. And the distribution is:



There is a very clear pattern that for Science students, Male students think they have better knowledge of the education choices and schedule that they should follow than Female students.

The target question “Final overall grade achieved in your secondary school” has p-value 0.2490993 of the chi-square fitness test, which means that the distribution has not much difference of male students and female students.

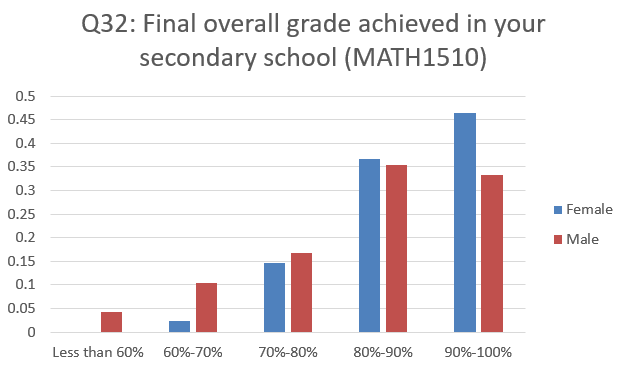


Now I separate the data from Class2 into Male group and Female group, and do the chi-square fitness test.

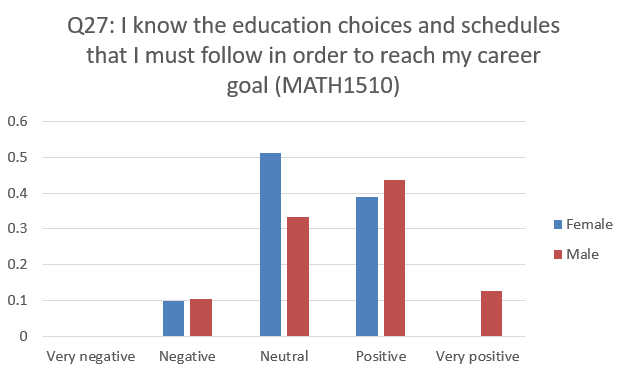
The p-value is generally larger than fitness test on Class1, which suggests that the distribution difference of features between gender in Class 1 is smaller than the distribution of gender in Class 2

|  |  |
| --- | --- |
| Question content | p-value of fitness test |
| I can do college level work successfully | 0.9507404 |
| I know that I am responsible for my own education | 0.9506711 |
| I schedule my study time carefully | 0.9363818 |
| I can make friends in new places | 0.9249588 |
| Excited about the courses I take | 0.9000241 |
| Raised hand in class | 0.8418236 |
| I do not need to study hard the night before an exam because of good planning | 0.8035867 |
| Getting a low-test score does not make me feel sad | 0.7919481 |
| School type | 0.744873 |
| I know a lot about myself in relation to what kind of profession would be best for me | 0.6857351 |
| I do not hesitate to ask for help | 0.6599229 |
| I am open to people with different opinions | 0.6547847 |
| Mother’s education | 0.5979913 |
| Available resources enough to meet my needs | 0.5749293 |
| I see education as something I will be doing throughout my life | 0.5500344 |
| Mother’s occupation | 0.4925968 |
| Father’s occupation | 0.4677013 |
| I write summary notes and I use them later when preparing for tests | 0.4586942 |
| I control my upset mood or anger without blaming others | 0.3445283 |
| Father’s education | 0.2878587 |
| Join MATH discussion group regularly | 0.245115 |
| Weekly study time in MATH on average | 0.2342807 |
| Reason you chose your former secondary school | 0.2188409 |
| Work on MATH problems via Internet resources | 0.2137499 |
| I know the education choices and schedules that I must follow in order to reach my career goal | 0.1834262 |
| My emotional health supports my ability to learn | 0.1420524 |
| I am able to manage time and other life demands effectively | 0.1346353 |
| Ask questions after class | 0.1255705 |
| I do outdoor indoor exercises regularly | 0.02906336 |
| I plan for each day of the week | 0.009562392 |
| Final overall grade achieved in your secondary school | 0.001227785 |

The target question “Final overall grade achieved in your secondary school” has the smallest p-value, but the intuitive distribution graph doesn’t vary that much. The difference is caused by the fact that there is more Male students in 60-70% while less students in 90-100%.



However, the p-value of the question “I know the education choices and schedules that I must follow in order to reach my career goal” has p-value 0.1834262 which is very large, the there exist the same distribution pattern as the Class 1 data.



We can make the conclusion that both in Class 1 and Class 2, Male students think they have better knowledge of the education choices and schedule that they should follow than Female students.

6 feature selection using genetic algorithm

## 6.1 genetic algorithm feature selection

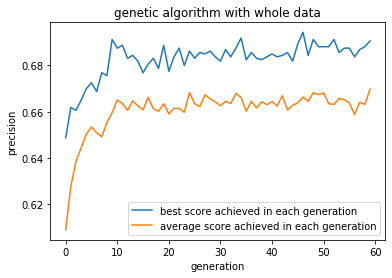
As pointed out in Chapter 4, the chi-square relationship test is not the best feature selection algorithm. In this chapter, genetic algorithm is used to selected the best feature combination.

Genetic algorithm is search heuristic that is inspired by Charles Darwin’s theory of natural evolution. Here I set the initial population as 40 and using Random Forest (Gini criterion and 200 trees) as the Fitness function. For each generation, I pick the top 5 genes with highest precision and then generate new population with crossover and mutation (with probability 0.05) on the selected 5 genes.

The precision of Fitness function is calculated from 5 times 10-folds-cv for each gene.

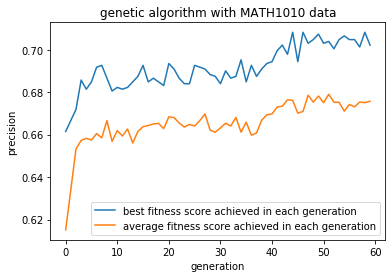
When applying genetic algorithm on whole data, the feature selected is :

1. Father’s education
2. Father’s occupation
3. School type
4. Reason you chose your former secondary school
5. Work on MATH problems via Internet resources
6. Raised hand in class
7. Ask questions after class
8. I schedule my study time carefully
9. Getting a low test score does not make me feel sad
10. Excited about the courses I take
11. I am open to people with different opinions
12. I plan for each day of the week
13. I know that I am responsible for my own education
14. I know a lot about myself in relation to what kind of profession would be best for me
15. I know the education choices and schedules that I must follow in order to reach my career goal
16. Available resources enough to meet my needs
17. I can do college-level work successfully
18. Class



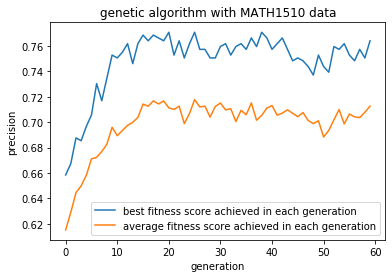
When applying genetic algorithm on Class 1 data, the feature selected is :

1. Father’s education
2. Father’s occupation
3. School type
4. Reason you chose your former secondary school
5. Weekly study time in MATH on average
6. I schedule my study time carefully
7. I am open to people with different opinions
8. I plan for each day of the week
9. I know the education choices and schedules that I must follow in order to reach my career goal
10. Available resources enough to meet my needs
11. I am able to manage time and other life demands effectively
12. I can do college-level work successfully



When applying genetic algorithm on Class 2 data, the feature selected is :

1. 'Mother’s occupation:',
2. 'Weekly study time in MATH on average:',
3. 'Work on MATH problems via Internet resources:',
4. 'Getting a low test score does not make me feel sad:',
5. 'I am open to people with different opinions:',
6. 'I know that I am responsible for my own education:',
7. 'I know a lot about myself in relation to what kind of profession would be best for me:',
8. 'Available resources enough to meet my needs:',
9. 'I do not hesitate to ask for help:',
10. 'I am able to manage time and other life demands effectively:',



And the precisions of Machine Learning methods trained using selected features on different data set are:

|  |  |  |  |
| --- | --- | --- | --- |
| Methods | Whole data | Class 1 data | Class 2 data |
| SVM | 0.6441874999999999 | 0.6465367965367969 | 0.7005617977528096 |
| Random Forest | 0.6659062499999997 | 0.6699567099567099 | 0.7384269662921351 |
| Random Forest + Adaboost | 0.6675937500000001 | 0.6695670995670995 | 0.7376404494382026 |

We can make the conclusion that by training model separately on two Classes, we can get better prediction precision. This result also inspires us that student background and interest is very important in predicted student learning outcomes.

Random Forest + Adaboost on whole data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| prediction  Reference | Less than 60% | 60%-70% | 70%-80% | 80%-90% | 90%-100% | precision |
| Less than 60% | 350 | 18 | 10 | 6 | 0 | 0.91146 |
| 60%-70% | 1 | 1731 | 39 | 43 | 110 | 0.89969 |
| 70%-80% | 691 | 839 | 5975 | 1374 | 719 | 0.6225 |
| 80%-90% | 386 | 81 | 1074 | 4762 | 926 | 0.6587 |
| 90%-100% | 272 | 731 | 1402 | 1915 | 8545 | 0.6642 |
| Recall | 0.20588 | 0.50912 | 0.7029 | 0.5879 | 0.8296 |  |

Random Forest on Class 1 data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| prediction  Reference | Less than 60% | 60%-70% | 70%-80% | 80%-90% | 90%-100% | precision |
| Less than 60% | 364 | 0 | 316 | 6 | 10 | 0.52299 |
| 60%-70% | 9 | 1559 | 78 | 19 | 31 | 0.91922 |
| 70%-80% | 809 | 467 | 5091 | 1555 | 680 | 0.5918 |
| 80%-90% | 40 | 225 | 761 | 2581 | 198 | 0.6783 |
| 90%-100% | 278 | 549 | 854 | 739 | 5881 | 0.7085 |
| Recall | 0.24267 | 0.55879 | 0.7170 | 0.5267 | 0.8649 |  |

Random Forest on Class 2 data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| prediction  Reference | Less than 60% | 60%-70% | 70%-80% | 80%-90% | 90%-100% | precision |
| Less than 60% | 0 | 0 | 0 | 0 | 0 | NaN |
| 60%-70% | 0 | 368 | 0 | 99 | 0 | 0.78801 |
| 70%-80% | 10 | 0 | 693 | 82 | 206 | 0.69929 |
| 80%-90% | 129 | 114 | 388 | 2631 | 414 | 0.7157 |
| 90%-100% | 61 | 118 | 319 | 388 | 2880 | 0.7647 |
| Recall | 0 | 0.61333 | 0.495 | 0.8222 | 0.8229 |  |

Surprisingly, although separate the data into 2 sets can improve the overall precision, this can also lead to bad predictor ability on “Less than 60%” predicting for both Class 1 and Class 2.

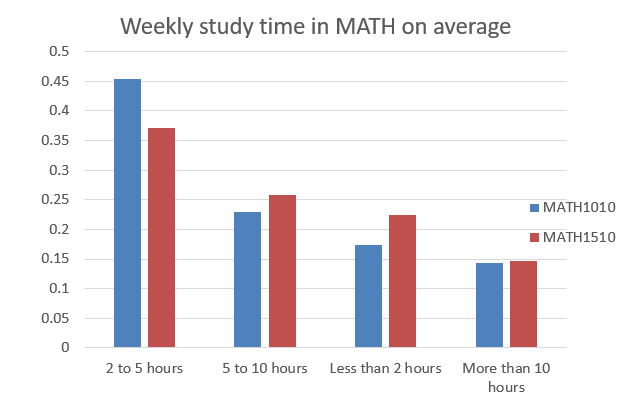
Larger dataset may be a suitable solution to solve this problem. And the detail will be discussed in next chapter.

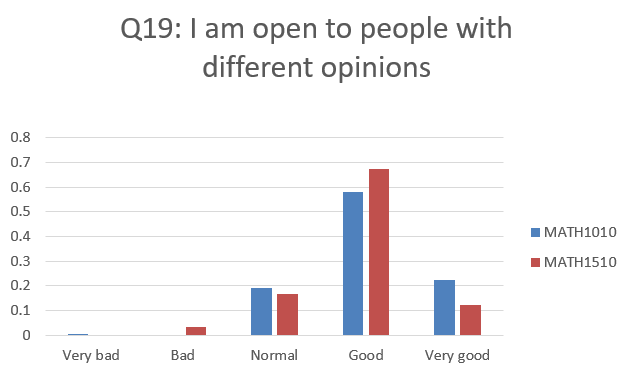
## 6.2 distribution of common feature selected in Class 1 & 2

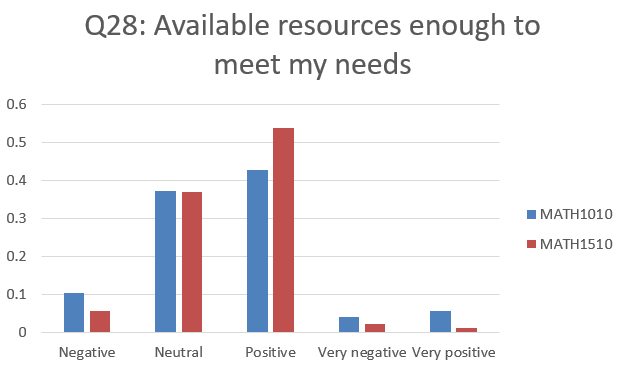
There are 4 features that shared by the selected features:

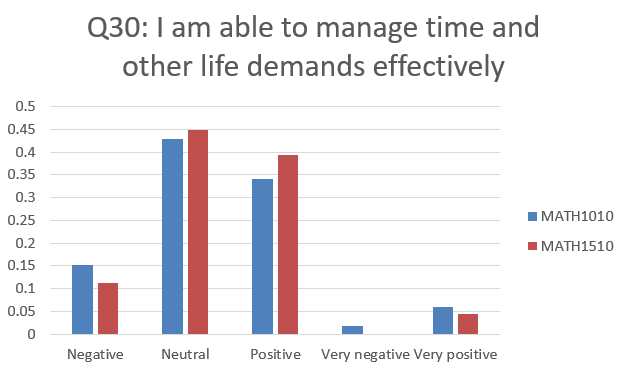
1. Weekly study time in MATH on average (fitness test p-value 0.3806002)
2. I am open to people with different opinions (fitness test p-value 0.0004499775)
3. Available resources enough to meet my needs (fitness test p-value 0.0802853)
4. I am able to manage time and other life demands effectively (fitness test p-value 0.4663642)

And their distribution of data from 2 Classes is:









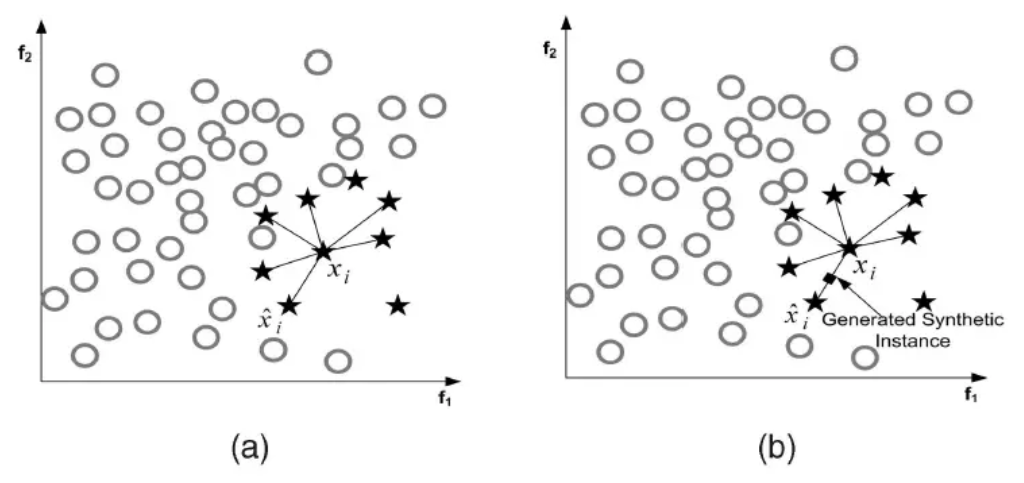
7 Imbalanced data handling

## 7.1 Up-sampling method

Generally speaking, imbalanced data will make the predictor biased. For example, if we our target variable is the Gender of people, while 98% instances in our sample is Male, then even the predictor just labels each input as “Male”, it still has 0.98 precision, which is not what we want.

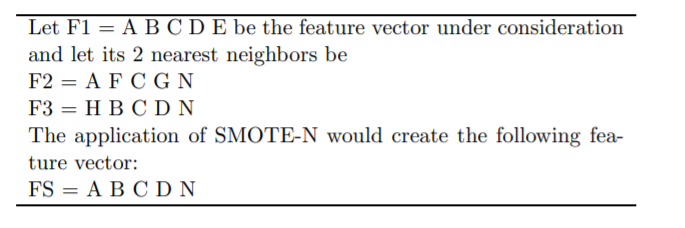
One common method to dealing with imbalanced data is up-sampling – simulated minority class data to make our set balanced. The advantages of up-sampling is that it can make data balanced as well as enlarge the data size a little bit. However, since up-sampling generates many synthetic instance, It may increasing the noise of our data set.

Synthetic Minority Oversampling Technique (SMOTE) is a common up-sampling way for numerical features. Its main idea is to simulate data of minority classes and make its size as large as majority class. For a real minority-class instance , find its k nearest neighbors. Random pick one point , and the simulated data is computed by the formula: (The convex combination of )



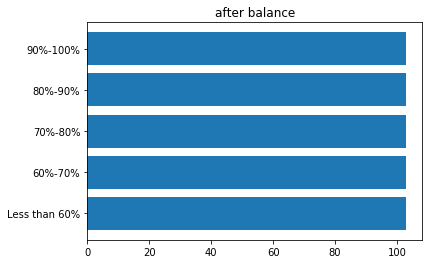
In this report, 5-nearst neighbor is used.

Synthetic Minority Oversampling Technique For Nominal variable (SMOTE-N) is a modification of SMOTE designed for categorical features. Unlike SMOTE, which using Euclidian distance to measure nearest neighbor, SMOTE-N uses Value Difference Metric (VDM), which looks at the overlap of feature values over all feature vectors, to measure the nearest neighbor, and the simulated data are generated by the pseudo code below:



Since Our data is originally categorical, but we constructed Machine Learning techniques on the numerical data set transformed by One-Hot-Encoding. Thus there both SMOTE-N followed by One-Hot-Encoding and One-Hot-Encoding followed by SMOTE will be tested:

The target variable distribution will look like:



I test the behavior of up-sampling algorithm on the whole data and Class 1 data (since Class 2 data size is very small), and the features used are the futures selected by genetic algorithm.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Without up-sampling | SMOTE | SMOTE-N |
| Random Forest + Adaboost (Whole data) | 0.6675937500000001 | 0.6542499999999998 | 0.6471875 |
| Random Forest (Class 1 data) | 0.6699567099567099 | 0.6532467532467531 | 0.6402597402597402 |

As the table suggests. The up-sampling method will make the precision decreasing because of the unreality of synthetic data.

## 7.2 precision orientated predictor

Numerical experiments suggest that up-sampling method will decrease the precision. However, the contingency table gives us new inspiration:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Random Forest + Adaboost (whole data) | Random Forest (Class 1 data) | Random Forest (Class 2 data) |
| Less than 60% | 0.91146 | 0.52299 | NaN |
| 60%-70% | 0.89969 | 0.91922 | 0.78801 |
| 70%-80% | 0.6225 | 0.5918 | 0.69929 |
| 80%-90% | 0.6587 | 0.6783 | 0.7157 |
| 90%-100% | 0.6642 | 0.7085 | 0.7647 |

The precisions for minority class “Less than 60%” are, as expected, very low, Which is 0.52299 and NaN (which means the predictor never labels data as “Less than 60%”) separately. However, when we using Random Forest + Adaboost on the whole data, the precision of “Less than 60%” is 0.92821.

Now let’s look at the precision of Random Forest + Adaboost which is trained on the whole data and tested using Class 1 and Class 2 seperately:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Random Forest + Adaboost trained on whole data and test on Class 1 | Random Forest + Adaboost trained on whole data and test on Class 2 | Random Forest trained on Class 1 data | Random Forest trained on Class 2 data |
| Less than 60% | 0.91146 | NaN | 0.52299 | NaN |
| 60%-70% | 0.88494 | 0.96021 | 0.91922 | 0.78801 |
| 70%-80% | 0.6418 | 0.50622 | 0.5918 | 0.69929 |
| 80%-90% | 0.6399 | 0.6805 | 0.6783 | 0.7157 |
| 90%-100% | 0.6466 | 0.7062 | 0.7085 | 0.7647 |

This observation inspires me to create a new “precision orientated predictor”, which combine all of the 3 predictors shown in above table:

1. Trained the Random Forest + Adaboost on the whole data using selected features, denoted as
2. Trained Forest on Class 1 & 2 data separately using selected features separately, and denoted them as
3. Tor any input instance x, predict its label by
4. If x belongs to Class 1, and x is labeled as 60%-70% or 80%-90% or 90%-100% (), then relabel x using

If x belongs to Class 2, and x is labeled as 70%-80% or 80%-90% or 90%-100% (), then relabel x using

After 100 times 10-folds-cv testing, the precision of predicting whole data can reach 0.7057812500000005,

Which is the highest precision score achieve in this report.

The contingency table is:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| prediction  Reference | Less than 60% | 60%-70% | 70%-80% | 80%-90% | 90%-100% | precision |
| Less than 60% | 356 | 10 | 106 | 2 | 22 | 0.71774 |
| 60%-70% | 1 | 1850 | 28 | 115 | 17 | 0.91994 |
| 70%-80% | 866 | 730 | 6651 | 1818 | 1163 | 0.5924 |
| 80%-90% | 208 | 829 | 5125 | 495 | 177 | 0.7499 |
| 90%-100% | 300 | 602 | 886 | 1040 | 8603 | 0.7526 |
| Recall | 0.20941 | 0.54412 | 0.7825 | 0.6327 | 0.8352 |  |

In conclusion, students background and interest (which affect whether they go to Science or Engineering department), are very important in predicting students learning outcomes.

8 graph method

There are totally 18 features selected to predict student learning outcomes for whole data:

1. Father’s education ·······················Q4
2. Father’s occupation ·····················Q5
3. School type ·························Q6
4. Reason you chose your former secondary school ··········Q7
5. Work on MATH problems via Internet resources ··········Q10
6. Raised hand in class ·····················Q11
7. Ask questions after class ····················Q12
8. I schedule my study time carefully ················Q14
9. Getting a low test score does not make me feel sad ·········Q15
10. Excited about the courses I take ·················Q16
11. I am open to people with different opinions ············Q19
12. I plan for each day of the week ·················Q22
13. I know that I am responsible for my own education ·········Q25
14. I know a lot about myself in relation to what kind of profession would be best for me ·························Q26
15. I know the education choices and schedules that I must follow in order to reach my career goal ·····················Q27
16. Available resources enough to meet my needs ···········Q28
17. I can do college-level work successfully ··············Q31
18. Class ···························Class